Semantics-Space-Time Cube. A Conceptual Framework for Systematic Analysis of Texts in Space and Time

Jie Li, Siming Chen, Wei Chen, Gennady Andrienko, and Natalia Andrienko

Abstract—We propose an approach to analyzing data in which texts are associated with spatial and temporal references with the aim to understand how the text semantics vary over space and time. To represent the semantics, we apply probabilistic topic modeling. After extracting a set of topics and representing the texts by vectors of topic weights, we aggregate the data into a data cube with the dimensions corresponding to the set of topics, the set of spatial locations (e.g., regions), and the time divided into suitable intervals according to the scale of the planned analysis. Each cube cell corresponds to a combination (topic, location, time interval) and contains aggregate measures characterizing the subset of the texts concerning this topic and having the spatial and temporal references within these location and interval. Based on this structure, we systematically describe the space of analysis tasks on exploring the interrelationships among the three heterogeneous information facets, semantics, space, and time. We introduce the operations of projecting and slicing the cube, which are used to decompose complex tasks into simpler subtasks. We then present a design of a visual analytics system intended to support these subtasks. To reduce the complexity of the user interface, we apply the principles of structural, visual, and operational uniformity while respecting the specific properties of each facet. The aggregated data are represented in three parallel views corresponding to the three facets and providing different complementary perspectives on the data. The views have similar look-and-feel to the extent allowed by the facet specifics. Uniform interactive operations applicable to any view support establishing links between the facets. The uniformity principle is also applied in supporting the projecting and slicing operations on the data cube. We evaluate the feasibility and utility of the approach by applying it in two analysis scenarios using geolocated social media data for studying people's reactions to social and natural events of different spatial and temporal scales.

Index Terms—spatiotemporal visualization, semantic visualization, data cube, interactive exploration, visual analytics.

1 Introduction

DATA cube [1] is a widely used metaphor representing organization of data along some dimensions of interest. For organizing and analyzing data that include texts with temporal and spatial references, such as geolocated social media posts (Fig. 1), we introduce a structure called Semantics-Space-Time Cube, or SSTC. In this structure, three dimensions correspond to (1) semantic categories, or topics, (2) locations (which may be regions in space), and (3) times (which may be time intervals). To understand how the text semantics varies over space and time, one needs to explore the complex and diverse relationships between the three heterogeneous information facets. Examples of such complex relationships are the spatial distribution of the text topics, temporal trend of topic popularity, spatio-temporal dynamics of topic appearance, etc. Our goal is to support the overall analysis task by visual analytics techniques that would not be too complex and difficult to use despite the complexity of the data and task.

For achieving this goal, we consider two problems. The first is how to construct the cube, in particular, how to represent text semantics in a summarized way suitable for being used as one of the cube dimensions. The second problem is how to utilize the cube structure and take advantage of it for supporting visual exploration of the data.

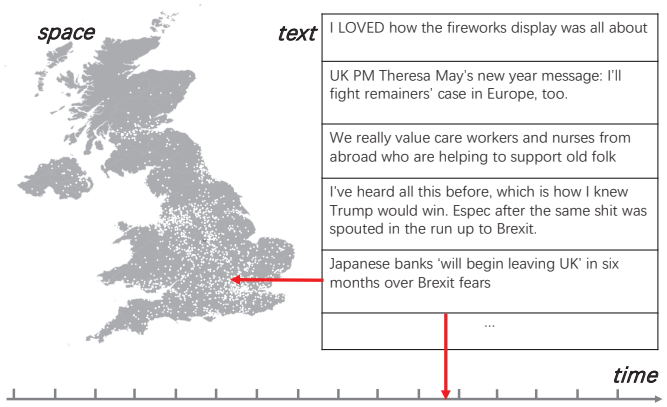


Fig. 1. Using geolocated Twitter data as an example to illustrate the target data structure. Each tweet is a text posted at some location and time moment.

A common way to summarize text semantics is probabilistic topic modeling [2]. However, the existing techniques do not work well on short texts. Besides, extracted topics may have ambiguous meanings or vary greatly when pa-

- *J. Li is with College of Intelligence and Computing, Tianjin University, China. jie.li@tju.edu.cn.*
- *S. Chen, G. Andrienko and N. Andrienko are with Fraunhofer Institute IAIS, Germany. S. Chen is also with University of Bonn, Germany. G. Andrienko and N. Andrienko are also with City University London. {siming.chen, gennady.andrienko, natalia.andrienko}@iais.fraunhofer.de*
- *Wei Chen is with State Key Lab of Cad&CG, ZheJiang University. chenwei@cad.zju.edu.cn.*

Manuscript received April 19, 2005; revised August 26, 2015.

rameters are slightly changed. Therefore, meaningful topics can hardly be extracted fully automatically, without human intervention. Our approach to solving these problems involves aggregation of short texts into larger documents and interactive selection of representative topics from results of multiple topic models using a visualization that reveals topic similarities and redundancies.

The problem of visual exploration of text-space-time data is challenging due to the high heterogeneity. The data components (texts, space, and time) differ extremely in their nature and properties. Such data cannot be treated as usual multidimensional data for which numerous analysis and visualization techniques exist. For example, in a parallel coordinates plot, multiple attributes are represented in a uniform way. In the case of heterogeneous structures, data components need to be visualized in different ways depending on their specific nature. While it is possible sometimes to show two components within one display (as in a space-time cube), this can hardly be done with three or more highly diverse components. One needs to use a combination of different displays and rely on interactive operations for uncovering and exploring relationships between them. Such a combination of visual and interactive tools is inevitably more complex than a single display and may be very difficult to use. A design challenge is to reduce the complexity and difficulty while providing sufficient functional power and enabling flexibility in exploration. To address this challenge, we, first, consistently utilize the SSTC metaphor and concepts of projection and slice, second, use similar organization principles for specific displays of the three diverse components and, third, propose a set of interaction operations uniformly applicable to each component.

Our main contributions are the following:

- A scheme of ***data transformation*** enabling systematic exploration of text semantics in space and time;
- Use of a ***cube*** metaphor for organizing data, defining the system of exploratory tasks, and designing the user interface and interactive operations;
- A ***workflow*** for analyzing text-space-time data, which involves characterization of text semantics through human-controlled topic modeling and interactive visual exploration of multifarious relationships between the text semantics, space, and time;
- A combination of ***visual and interactive tools*** that support the variety of exploratory tasks. The tool design implements the principles of structural, visual, and operational uniformity for reducing the UI complexity.

The paper is structured as follows. After introducing the problem statement (Section 2), we review the related work (Section 3), present our approach (Section 4), describe and substantiate our visual design (Section 5), and demonstrate the application of the approach in two case studies (Section 6). Expert feedback is presented in Section 7, followed by a discussion and conclusion in Section 8.

2 Problem Statement

We describe the structure of the data we are dealing with, introduce the cube metaphor, and define the system of tasks on exploring relationships between text semantics, space, and time.

2.1 Data

The original data format is $< text, location, time >$. Locations can be specified by coordinates or names of geographical places. Texts are arbitrary and unstructured. It is necessary for analysis to represent text semantics in a structured way. We assume that text semantics can be characterized using a finite set of *topics* (themes) of interest. For each text, it can be determined how much it is related to each topic. The degree or the likelihood of relatedness can be expressed numerically, e.g., as a number between 0 (unrelated) to 1 (uniquely related). We shall call this number *topic weight*. Hence, each text is represented by a vector of topic weights $TWV =< w_1, w_2, ..., w_N >$, where N is the number of topics considered. The data structure is thus transformed to $< TWV, location, time >$.

2.2 Cube

We aggregate and organize the data so transformed along three dimensions comprising the topics, locations, and times, denoted P (to$P$ics), S (Space), and T (Time), respectively. The Cartesian product $P \times S \times T$ is metaphorically called *semantics-space-time cube*. For practical purposes, all three sets P, S, and T need to be discrete and finite. The set of topics P is discrete and finite by construction. The space and time are discretized by partitioning into suitable regions and time intervals, respectively. Any combination (p, s, t) composed of a particular topic $p \in P$, location (region) $s \in S$, and time interval (also called *time step*) $t \in T$ will be called a *point* of the cube. For each cube point (p, s, t), we derive several *measures* from the data, which include

- the *popularity* score of the topic p at location s during time t, which is calculated as the mean weight of this topic in the messages that were posted at location s during time t [3];
- a keyword vector consisting of pairs <keyword, weight>, where *weight* is a numeric measure that can represent the importance of the keyword in the topic p at location s during time t.

The resulting data structure is $P \times S \times T \rightarrow (PS, KW)$, where PS and KW stand for the *P*opularity *S*core and the *K*eyword *W*eight vector, respectively.

2.3 Slices and Projections

The overall analysis task is to explore the variations of the popularity and keyword usage along the three dimensions of the cube. However, the high dimensionality of the structure $P \times S \times T$ does not allow seeing the entire variation, which requires the overall task to be decomposed into simpler subtasks. The whole variation can be viewed as a function of multiple variables and the analysis task as the task of studying the behavior of this function [4]. The overall behavior can be studied by considering its *slices*, in which the value of one variable is fixed for exploring the variation over the remaining variables. Using the cube metaphor, a behavior slice corresponds a cutting plane in the cube that is parallel to one of its faces (Fig. 2a-c).

Multiple slices corresponding to different values of one variable can be aggregated by putting these values together and applying some summarizing operators (sum, mean,

mode, quartiles, etc.) to the respective values of the function. This is done for each combination of the values of the other variables. The result is called a *projection*, because it can be metaphorically seen as a projection of the cube content onto one of its faces (Fig. 2d-f). The data structure of a projection is the same as in the slices from which it was obtained whereas the values within this structure are aggregates of the values from the slices.

According to the functional view of data and tasks [4], we shall use a formal notation based on the representation of the cube as $P \times S \times T \rightarrow (PS, KW)$, which also represents the overall analysis task: how PS and KW vary over the whole set $P \times S \times T$. The measures PS and KW are independent of each other, the variation of each of them can be explored separately, resulting in two subtasks $P \times S \times T \rightarrow PS$ and $P \times S \times T \rightarrow KW$. In the following, we shall consider both subtasks in a generic way using the notation $P \times S \times T \rightarrow M$, where M represents PS, KW, or, generally, any other meaningful measure that can be defined and calculated from the data. Fore example, one may apply sentiment analysis [5] to calculate the fractions of the texts with positive, neutral, and negative expressions concerning the topics.

The *selection* operation takes a *slice* in the cube corresponding to one selected value within one dimension (Fig. 2a-c). The corresponding analysis tasks are thus conducted within the slice to address the variation of the studied measure over the "plane" formed by the combinations of the values from the two other dimensions. The possible types of analysis tasks based on cube slices are:

- $t \rightarrow (P \times S \rightarrow M)$ - for a selected time step $t \in T$, study the commonalities and differences among the spatial distributions of the measures of different topics.
- $s \rightarrow (P \times T \rightarrow M)$ - for a selected location $s \in S$, study the measures of different topics and their changes over time;
- $p \rightarrow (S \times T \rightarrow M)$ - for a selected topic $p \in P$, study the variation of the measure over space and time;

The *aggregation* operation creates a *projection* of the cube content along one dimension (Fig. 2d-f). Aggregation can be applied not only to the set P, S, or T as a whole but to a subset of it. Possible analysis tasks address the variation of the summary values over the projection plane. The task types can be defined analogously to the slice-based tasks:

- $\Sigma(T) \rightarrow (P \times S \rightarrow M)$ - for all times taken together, study the commonalities and differences among the spatial distributions of the aggregate measures of different topics.
- $\Sigma(S) \rightarrow (P \times T \rightarrow M)$ - for all locations taken together, study the variation of the aggregate measures of different topics over time;
- $\Sigma(P) \rightarrow (S \times T \rightarrow M)$ - for all topics taken together, study the variation of the aggregate measures over space and time;

2.4 Analysis tasks

Both slice-based and projection-based tasks address the variation of some measure over a "plane", i.e., a Cartesian product of two dimensions. Let us represent such variation in a general way as a function $X \times Y \rightarrow Z$. It can be explored in two complementary ways [4]:

- Consider the variation of the function $Y \rightarrow Z$ over the set X, represented as $X \rightarrow Y \rightarrow Z$. Thus, for $S \times T \rightarrow PS$, it is "consider how the temporal evolution of the topic popularity varies over the space": $S \rightarrow T \rightarrow PS$.
- Consider the variation of the function $X \rightarrow Z$ over the set Y, represented as $Y \rightarrow X \rightarrow Z$. Taking the example of $S \times T \rightarrow PS$, it is "consider how the spatial distribution of the topic popularity changes over time": $T \rightarrow S \rightarrow PS$.

Using this general schema and treating slice- and aggregation-based tasks in a uniform way, we define the following system of task types:

- Tasks based on time slices $t \rightarrow (P \times S \rightarrow M)$ or time projection $\Sigma(T) \rightarrow (P \times S \rightarrow M)$:
  - **T1**: $S \rightarrow P \rightarrow M$, e.g., which topics were popular at different locations
  - **T2**: $P \rightarrow S \rightarrow M$, e.g., where did different topics receive more attention
- Tasks based on space slices $s \rightarrow (P \times T \rightarrow M)$ or space projection $\Sigma(S) \rightarrow (P \times T \rightarrow M)$:
  - **T3**: $P \rightarrow T \rightarrow M$, e.g., when and how long were different topics popular
  - **T4**: $T \rightarrow P \rightarrow M$, e.g., how did the relative popularities of the topics differ among the times
- Tasks based on topic slices $p \rightarrow (S \times T \rightarrow M)$ or topic projection $\Sigma(P) \rightarrow (S \times T \rightarrow M)$:
  - **T5**: $S \rightarrow T \rightarrow M$, e.g., what are the temporal variation trends at different locations
  - **T6**: $T \rightarrow S \rightarrow M$, e.g., how does the spatial distribution of the measure change from time to time

Furthermore, there may be tasks on *comparing* the variations in two or more slices. Hence, each task type T1-T6 in application to slices can be split into two subtypes, *behavior characterization* and *behavior comparison* [4]. To keep the presentation simpler, we shall not introduce additional formal representations for slice comparison tasks, but it should be kept in mind that the task types T1-T6 cover both characterization and comparison tasks.

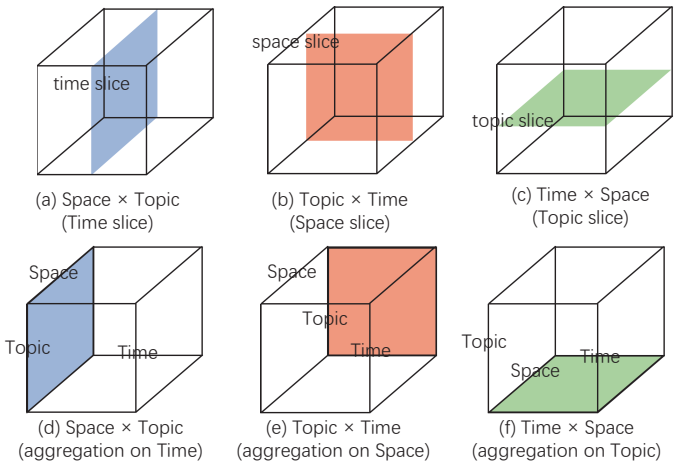


Fig. 2. Results of selection and aggregation operations are symbolically represented as slices (a-c) and projections (d-f) of the cube.

2.5 Design Goal

Our goal is to design and implement such an approach to supporting all the defined tasks that all the tasks could be fulfilled in a possibly uniform way. The reason for striving towards the uniformity is a desire to make the whole exploration process easier for the analyst. Since the overall analysis task has to be decomposed into multiple diverse subtasks, a possibility to perform these subtasks in similar ways can reduce the cognitive effort required for learning and remembering the functionality provided.

3 Related Work

The related research includes works in visual analytics and cognate disciplines that (1) apply topic modeling to represent text semantics, or (2) deal with data having textual, spatial, and temporal components, or (3) utilize different variants of data cubes for supporting data analysis, or (4) use cubes to represent data visually. This section is structured according to these four themes.

3.1 Topic Modeling in Visual Analytics

Probabilistic topic modeling [6] is a class of natural language processing methods that are used for obtaining a structured representation of collections of texts. From a given text collection, these methods extract a set of of latent topics, where each topic is a probability distribution over words of a chosen base vocabulary. The texts are represented by vectors of topic probabilities, or weights [2]. Examples of topic modeling methods are Latent Dirichlet Allocation (LDA) [2] and non-negative matrix factorization [7].

Topic modeling has been widely used in visual analytics works focusing on text analysis [8], [9], [10]. Texts from social media have received much attention [11], [12]. In Recompile 3 particular, researchers have focused on evolution of topics and their relationships. Xu et al. [13] proposed an topic competition model to characterize the competition for public attention on multiple topics. Sun et al. [14] extended this model for analyzing both competition and cooperation relationships among different topics in social media. Wang et al. [15] proposed a visual analytics system for analyzing the topic transmission between different social groups. Dou et al. [16] supported exploration of hierarchical relationships among topics. Cui et al. [17] focused on dynamic hierarchical relationship among topics in different time periods.

While a large body of research has been done on analyzing text topics over time, we are not aware of works that would consider the relationships of topics to a more complex spatio-temporal context. Our work makes the first step in this direction.

3.2 Exploration of Texts in Space and Time

ThemeRiver [18] is perhaps the best known technique to visualize changes of text semantics over time. Themes extracted from texts are represented along a time axis by bands ("currents") with the widths proportional to the topic strengths. Since the first publication, the idea was actively used and adapted to a variety of tasks, such as analysis of opinion diffusion [19] and anomalous information spreading [20]. Liu et al. [21] combine the metaphors of river and sedimentation to show older data in aggregated form and more recent in detail. Since ThemeRiver has proved its effectiveness and gained high popularity, we used this idea in our time view display.

In the visual analytics research dealing with spatially and temporally referenced text data, such as geolocated social media posts, many works have been focusing on keywords occurring in the texts. The simplest approach is to extract a subset of data containing occurrences of specific keywords and analyze the spatio-temporal distribution of the selected data, particularly, to detect spatio-temporal clusters [22], [23], [24]. In these works, interactive visual analysis is applied only to the spatial and temporal aspects of the data.

Another approach is to process the texts for detecting references to *events*, i.e., occurrences at specific times and places. Events are identified from groups of texts mentioning the same places and same or overlapping times [25]. Markus et al. [26] detect events from peaks of high tweeting activity and meaningfully label them using keywords from the tweets. Zhou and Xu [27] identify events using bursty word detection techniques from machine learning. Chae et al. [28] identify abnormal events using seasonal-trend decomposition. After extracting the events, their relationships to space and time are explored visually using map- and time line- or calendar-based displays whereas the texts related to the events are summarized into word clouds [29], [30], [31]. Additionally, the sentiments of event-related texts can be visually explored [26], [27].

Bosh et al. [32] proposed a system ScatterBlogs for visual detection of events based on multiple occurrences of the same keyword in messages posted at nearby places and times. Such keywords are shown on a map at the places where they occurred using font sizes proportional to the number of occurrences. ScatterBlogs2 [33] extends ScatterBlogs by adding sophisticated tools for text filtering, so that the visual analysis is applied to previously selected potentially relevant texts, such as texts mentioning natural disasters. Chae et al. [34] propose tools for analyzing the spatio-temporal distribution of Twitter users based on the locations and times of the posted tweets. For a detected spatio-temporal cluster of Twitter users, the analyst can run topic modeling on the posted tweets and see the extracted topics represented by a word cloud, which can provide a hint concerning the event that caused the people to convene.

A common feature of all these works is their focus on exploring extracted events rather than texts with their semantics. A different focus is taken by Cao et al. [35]: they propose a visualization that shows re-postings of messages mentioning specific events for exploring the diffusion of information through social media. However, like the other works, this work does not address the variation of text semantics over space and time.

There was a case study in which the spatial and temporal distributions of text topics were explored separately [36]. The researchers created a predefined set of topics, such as 'work', 'transport', 'food', 'sport', etc., and specified a list of relevant keywords for each topic. Twitter posts were associated with the topics through detecting keyword occurrences. Based on the posting times, the researchers explored the distribution of each topic over the daily and weekly

time cycles; based on the tweet locations, they explored the distribution of the topics over the territory of a city. The research goal was to investigate to what extent the social media reflect people's current activities, which is an example of a specific analysis task. The researchers neither intended to define the full space of tasks nor proposed a system or framework for systematic support of the tasks.

3.3 Data Cube

Data cube [1] is a model for organizing multidimensional data designed to support OLAP (On-line Analytical Processing) queries. For constructing a cube, some data fields (attributes) are chosen as *dimensions* and others as *measures*. For each combination of values of the dimensions, the cube contains corresponding measures. In OLAP, hierarchical aggregation is applied to the dimensions, and the corresponding aggregated measures, such as sum, $average$, $count$, etc., are pre-calculated and stored. The model has been used and adapted to represent datasets in different domains, such as social media [37], traffic [38], and graph analysis [39]. More recently, several approaches, such as NanoCubes [40], Hashedcubes [41], Gaussian Cubes [42] and Time Lattice [43], have been proposed for performing specific types of tasks on large datasets. These works focus on optimizing the data structure to enable real-time response to interactive operations, whereas our focus is supporting the exploration of various relationships among dimensions.

Most commonly, cells in a data cube contain numeric values. In Text Cube [44], which aims to support OLAP queries on multidimensional datasets with text fields, the cells contain unstructured texts. Zhang et al. [45] construct a Topic Cube using a hierarchy of topics obtained through text analysis. The topics form one of the cube dimensions while the others may consist of spatial locations, dates, and times (day or night). The cube cells contain two kinds of measures, namely, word distribution of a topic and topic coverage by documents. This structure is similar to what we use. While Zhang et al. focus on efficient construction of the topic cube and support of OLAP queries, our focus is enabling visual exploration of relationships between topics, space, and time.

3.4 Space-Time Cube as a Concept and a Visualization Technique

The idea of SSTC resembles the concept of space-time cube (STC), which was introduced by T. Hägerstrand [46] in late 60s as a metaphor for representing human behavior in geographic space and time. The space is represented by two dimensions of the cube and time by the third dimension. This idea was implemented as a visualization technique for exploring the spatio-temporal distribution of discrete objects, such as events [47] and trajectories [48], [49]. Bach et al. published a review of visualization applications based on STC [50]. Amini et al. [51] conducted an experiment on the use of interactive 2D and 3D displays of trajectories and found that the 3D display outperformed the 2D one for some tasks and was also liked more by the participants.

The STC technique can work well for representing objects that are either sparsely spread within the cube or grouped in clusters, which collectively occupy only a small fraction of the cube volume. In other cases, distribution patterns can hardly be identified due to occlusion. Dübel et al. [52] systematically discuss the advantages and disadvantages of 2D and 3D visualizations of spatial and spatio-temporal data. General disadvantages of 3D displays are occlusion, distortion, and difficulty of matching objects to their spatial locations, but 3D views may still be good for exhibiting clusters and some other kinds of distribution.

There were also attempts to use STC for representing spatial time series (e.g., [53, pp.107-108]), i.e., data in which some attribute value exists for each combination of location and time step. Such data can be shown reasonably well in an STC only when there are relatively few distinct locations, the time series are not very long, and small attribute values are hidden. Even under these conditions, the display involves much occlusion, and it may become fully ineffective when these conditions do not hold.

The SSTC differs from the STC by including a semantic dimension composed of topics. As a data structure, SSTC does not reflect the inherent dimensionality of the spatial component, the space being represented as a set of discrete locations, which forms a single dimension of the cube. However, this does not imply that the visual representation of the spatial component must also be unidimensional. On the opposite, the visual representation must respect the inherent properties of the spatial component and look familiar and understandable to the user. This means that the locations need to be represented on a map. Consequently, we cannot create a single display representing simultaneously the spatial, temporal, and thematic dimensions of the SSTC. Instead, we must use a combination of partial views, which, in principle, may be 2D or 3D. The disadvantages of 3D views [52] are very relevant to our data structure, in which, similarly to spatial time series, a measure exists for each combination of location, time step, and topic; hence, the cube is fully filled with values. Therefore, our logical choice is 2D displays.

4 Approach Overview

4.1 Workflow

We propose an analytic workflow presented in Fig. 3. The original data available in the form <text, location, time> (Fig. 3a) are transformed, as explained in Section 2.1, into the form <TWV, location, time> suitable for cube construction. The transformation (Fig. 3b) involves definition of discrete finite sets of topics, locations, and time steps. The set of topics is obtained using the LDA (Latent Dirichlet Allocation) method applied after basic text preprocessing including stop-word elimination and lemmatization. The sets of locations and time steps are defined through discretization of space and time. The next step is construction of the cube (Fig. 3c), which involves calculation of the topic popularity and keyword weights for each point of the cube (see Section 2.4). The resulting cube is then explored using visual and interactive techniques (Fig. 3d).

4.2 Cube Creation

Cube creation involves definition of the cube dimensions, i.e., the sets of topics, locations, and time steps, and calculation of the measures for each point in the cube (see Sections 2.1 and 2.4).

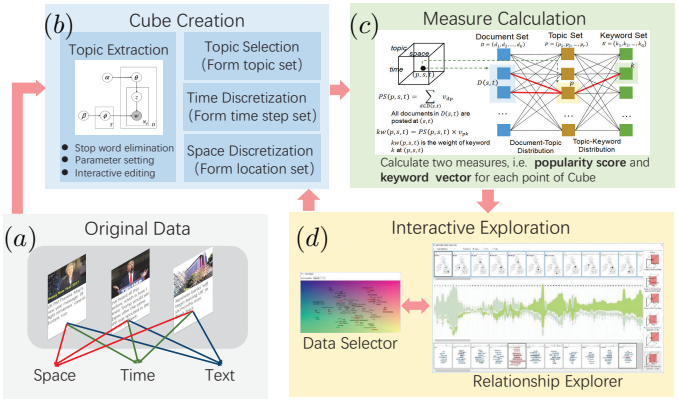


Fig. 3. Data processing and analysis workflow.

TABLE 1
An example of LDA result.

index	keywords
1	scotland, deal, brexiteers, indyref, means, talks
2	economy, money, pound, due, market, vote, property
3	euref, leave, vote, referendum, remain, eureferendum
4	nhs, love, peace, france, freedom, irexit, frexitl
5	brexitshambles, article, theresamay, nobrexit, court
6	remain, brexitbritain, lies, brexiters, campaign
7	post, postbrexit, impact, trade, future, great
8	trump, world, people, time, america, election, wrong
9	stopbrexit, manchester, march, grexit, exitfrombrexit
10	ireland, english, trndnl, gmt, brexitbill, northern
11	bbc, read, latest, post, blog, blair, interview
12	hate, people, britain, brexiteers, price, racist

4.2.1 Topic Extraction

We employ Latent Dirichlet Allocation (LDA) [2], a probabilistic topic modeling method that is often used to summarize vast amounts of texts. A topic is defined as a probability distribution over a given vocabulary, i.e., a set of keywords, where keywords with high probabilities represent the semantic content of the topic. Apart from the $topic - keywords$ distributions, an LDA model also produces $document - topics$ distributions consisting of the probabilities of each document to be related to each topic.

The LDA model does not work well on short texts [54]. A simple but popular way to alleviate the problem is to aggregate short texts into longer pseudo-documents [55], [56], [57]. The texts that are merged into a single document should be semantically related. Texts posted in social media usually contain hashtags, and the use of the same hashtags can be treated as indication of semantic relatedness of the texts. However, the meanings associated with the hashtags may change over time. We therefore aggregate messages that not only have common hashtags but also have close times of posting. The effectiveness of this approach has been proven in our previous work [3].

Table 1 shows representative keywords of topics extracted from a dataset of tweets related to Brexit that were posted in Britain from March 2016 to October 2017. During that period, the Brexit was a hot discussion topic in the British society. The keywords in each topic are sorted according to their weights (i.e., probabilities) for the topics. Since each list begins with a different keyword, these most important keywords can be used as topic representatives.

Since we apply the LDA method to pseudo-documents obtained by aggregation of the original documents (i.e., short messages), it generates the topic probability distributions for the pseudo-documents. We propagate these distributions to the original documents in the following way. If a document has been included in a single pseudo-document, it receives the topic probability distribution of this pseudo-document. If a document has been included in several pseudo-documents, the corresponding topic distribution vector is composed of the mean probabilities of the topics computed from the probabilities for the pseudo-documents.

4.2.2 Topic selection

The LDA and other topic modeling algorithms require setting the number of topics to generate, which is a parameter of the algorithms. It may be very hard to estimate how many meaningful and distinct topics may exist in a text collection. The choice of the parameter value may have high impact on the result. Some topics can only be found with specific values of the parameter, and even a slight change of the value may lead to extracting a very different set of topics [28].

To reduce the effects of the parameter value choice and be able to compare topics generated with different parameter setting, we utilize an ensemble-based approach. We run LDA multiple times with different parameter values, e.g., $n \in \{10, 20, 30, 40\}$. We then create a visual display by projecting all extracted topics on a 2D plane according to their keyword probability distributions by means of a dimensionality reduction method, such as t-SNE [58] or MDS [59]. The topics are represented by points labeled by the keywords with the highest probabilities. Longer keyword lists, as in Table 1, are shown upon mouse-hovering. In this display, very similar topics resulting from different runs will have very close positions.

The topic projection display is used for selecting a topic set for the cube construction. The selected topics should be semantically diverse; therefore, from a group of close topics, it is sufficient to pick a single topic. Furthermore, the analyst may find some topics uninteresting, or vague, or irrelevant to the analysis goals; such topics do not need to be included in the cube.

4.2.3 Space and time discretization

Generally, the spatial and temporal domains are continuous. Since cube construction requires discrete sets of locations and time steps, the space and time need to be discretized. In some applications, predefined space divisions can be suitable, such as administrative division into countries, provinces, cities, etc. Other possibilities for space discretization include regular division by rectangular or hexagonal grid or irregular tessellation based on the spatial distribution of the data [53]. In studies of human mobility behavior, it is often appropriate to use individual and public activity locations that can be extracted from long-term data [60].

The temporal domain is partitioned into suitable intervals. It is desirable to define the intervals based on calendar units, such as months, weeks, or days, or, when a finer temporal granularity is needed, hours of the day. Such intervals are meaningful to humans. For example, in Fig.

6, the location set consists of the important cites of Britain, and the time period is divided into weeks.

It is possible to use a sliding window along the time axis for considering overlapping time intervals, for example, intervals of the length 1 week with a shift of 1 day.

4.3 Measure Calculation

We calculate the popularity score and keyword weight vector, as defined in Section 2.4, for each point (p, s, t) in the cube. For this purpose, we utilize the $document - topic$ and the $topic - keyword$ distribution outputs of the LDA model, as shown in Fig. 4.

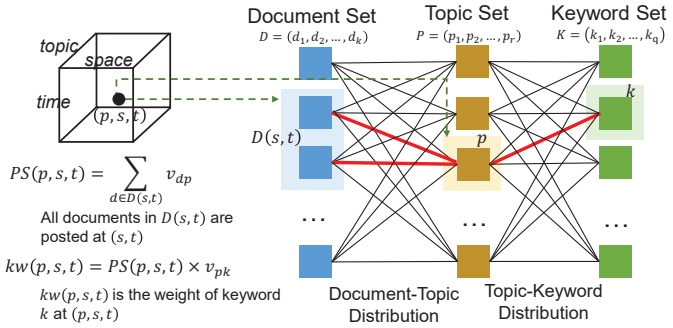


Fig. 4. Illustration of the calculations of the popularity score and the keyword weight vector.

In the $document - topics$ distribution, the sets of documents and topics form a bipartite graph, where each edge represents the probability of a document being related to a topic. Similarly, the $topic - keywords$ distribution can be seen as a bipartite graph where the edges represent the keyword weights for the topics, as in Fig. 4b.

To calculate the popularity score $PS(p, s, t)$ for the point (p, s, t), we collect all documents that were posted at s and t, thus forming a document subset $D(s, t)$, and calculate the sum of their weights on the topic p. Using Fig. 4b as an example, two documents are posted at s and t (marked with light blue shade), thus the popularity score for the point $PS(p, s, t)$ is the sum of the weights of the two documents on the topic p (the red lines). We multiply $PS(p, s, t)$ by the weight of each keyword for the topic p (obtained from the LDA model) to obtain $kw(p, s, t)$, which represents the prominence of keyword k at the point (p, s, t), as in Fig. 4a. If a topic has a high popularity at a point and a keyword has a high weight for the topic, the prominence (weight) of the keyword at this point is high.

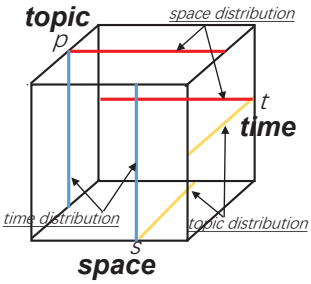
4.4 Interactive Exploration

The interactive exploration of the data organized in the cube is supported by two components, ***Data Selector*** and ***Relationship Explorer***. Their functions are described below whereas the visual design is described in detail in the following section.

4.4.1 Data Selector

The analyst may not necessarily need to deal with the whole cube at each moment of the analysis. Only a subset of the topics, or locations, or time intervals may be relevant to the current analysis focus. The Data Selector supports the selection of relevant subsets of topics, locations, and/or times. This defines a sub-cube of the whole cube. The subsequent exploration is done on the sub-cube; the remaining part of the cube is not reflected in the visual displays. The analyst can modify the selection at any moment in the process of analysis, when it is necessary to consider another subset of the data.

To define a sub-cube for further exploration, the analyst separately selects subsets of topics, locations, and times. The selection is supported by an interactive visual display that informs the analyst about the similarities between the elements of the currently considered dimension. This is done by projecting the elements onto a 2D plane according to their similarities by means of an appropriate multidimensional reduction algorithm, such as t-SNE [58] or MDS [59]. The approach is the same as is used for the initial topic set selection (Section 4.2.2); however, different feature vectors are used. For the initial topic selection, the feature vectors were the keyword probability distributions. In this case, the feature vectors are constructed based on the cube structure. For an element of the currently considered dimension, the feature vector may be composed from the values lying within the corresponding slice (Fig. 2a-c). Such a vector represents the distribution of the measures corresponding to the element over the other two dimensions. Another possibility is to define a feature vector based on the projection of the cube content along one of the two other dimensions (Fig. 2d-f). Such a vector represents the distribution of the aggregated measures over the remaining dimension.



topic	k_1	...	k_m
p_1	v_{11}	...	v_{1m}
p_2	v_{21}	...	v_{2m}
...	...	...	...
p_n	v_{n1}	...	v_{nm}

keyword space

(b)

Fig. 5. Feature vector selection. (a) Each object, i.e. location, time step and topic, can be projected according to the distribution of the corresponding measures over one or two other dimensions. (b) The feature vectors that are used for the initial topic selection (Section 4.2.2).

Hence, the topics can be laid out in a projection based on the distributions of the corresponding measures over time and/or space, the locations can be laid out based on the distributions of the corresponding measures over topics and/or time intervals, and the time intervals can be laid out according to the distributions of the corresponding measures over the topics and/or locations. Figure 5a illustrates the construction of the feature vectors, and Figure 5b shows the structure of the feature vectors used for the initial topic selection.

4.4.2 Relationship Explorer

The Relationship Explorer enables exploration of the data contained in the selected sub-cube. It represents the data distributions along the three cube dimensions, space, time,

and topics. Interactive operations enable the analyst to see various relationships existing among the distributions. The component is described in detail in the following section.

5 Visual Design

In designing the visual interface (Fig. 6), we strove to restrict its complexity and difficulty for learning and use notwithstanding the complex structure of the data dealt with. To moderate the complexity, we developed an idea of design uniformity, which is explained below.

5.1 Design Uniformity Principle

The visual interface needs to represent three heterogeneous dimensions of the data, space, time, and topics, and the corresponding measures, i.e., topic popularity and keyword weight vector. The representation of the dimensions needs to reflect their inherent properties, particularly, the geographic arrangement of the locations and the linear ordering of the time intervals. As discussed in Section 3.4, all three dimensions cannot be suitably represented in a single view. We have to use several complementary views, and 2D representation is preferred over 3D. Based on these premises, we come to the necessity of using three views, spatial, temporal, and topical. The first two reflect the inherent properties of the space and time, respectively, and the third reflects the composition of the topical dimension from different topics.

The idea of uniformity means similar appearance of the three views and similar ways of interacting with them. The implementation of this idea involves several aspects as discussed below.

***Structure Uniformity* (SU)**. We want the views to have a similar structure, i.e., similar layouts of display elements or components. The temporal view has to have a linear layout for reflecting the inherent linear ordering among the time intervals. For consistency, we adopt a linear layout also for the other two components. Specifically, we represent topics in the topical view by linearly arranged components, called 'cards', showing topic-specific information. Correspondingly, the spatial view includes multiple linearly arranged cards, which contain geographic maps to reflect the specifics of the spatial facet. Unlike the other two facets, time is a continuous succession of time intervals. This inherent feature is reflected in representing data by continuous curves or shapes, which is unique for the temporal view. Hence, the views are structurally similar to the extent allowed by the inherent properties of the data facets they represent, but they also have unique features reflecting the specifics of these facets.

***Visual Style Uniformity* (VU)**. The use of colors, fonts, labeling, highlighting, and other kinds of visual marks and variables need to be consistent between the views. Similar elements must have similar meanings irrespective of the views in which they appear.

***Interaction Uniformity* (IU)**. Since the representation of the data is decomposed into three views, interactive operations are required for establishing links among the views and exploring the distribution of the measures across the dimensions. The same interactive techniques need to be available in all views, and their implementation must be consistent between the views.

***Operation Uniformity* (OU)**. As discussed in Section 2.3, the task of studying the overall data distribution is decomposed into simpler subtasks dealing with data cube slices and projections. In the process of exploration, the analyst should be able to choose slices and projections to consider next. The analyst may also wish to restrict the further exploration to a sub-cube of the data (Section 4.4.1). In supporting the selection of sub-cubes, slices, and projections, the cube dimensions should be treated uniformly. Furthermore, the choice of a perspective on the data (i.e., which dimension is fixed or aggregated while the others preserve variation) should not affect the structure and visual appearance of the spatial, temporal, and topical views.

The use of the cube metaphor, in which space, time, and topics are treated as uniform data dimensions, provides a good basis for designing the UI according to the uniformity principle. In the following, we describe the resulting design.

5.2 Data Selector

The Data Selector enables selection of data sub-cubes for further interactive exploration (Section 4.4.1). The controls for sub-cube selection do not need to be present on the screen constantly. They appear in a pop-up window triggered by clicking on a special button.

The Data Selector (Fig. 6g) consists of three panels that are integrated in a tab-control. The panels correspond to the three dimensions of the cube. In each panel, the elements of the respective dimension are arranged in a 2D layout reflecting the similarities between the corresponding distributions of the data over one or two of the remaining dimensions, as described in Section 4.4.1. The dimensions to use are selected through a drop-down list. The items shown in the projection are labeled, depending on their nature, with the location names, indexes of time steps, or dominant keywords (i.e., having the highest weights for the topics). Collision detection and resolution are utilized to avoid overlaps of the labels. Currently selected items are highlighted. The projection background is colored using a continuous two-dimensional color scale. The purpose of the coloring is explained below.

Color Assignment. In the Relationship Explorer, we consistently use colors for representing the same data items in different views. An arbitrary assignment of colors to elements of a data dimension would result in using too many unrelated and uninterpretable colors, which complicated and impedes perception and cognition due to the limited human attention capacity [61]. To deal with this problem, colors need to be assigned in a meaningful way. Following the approach proposed by Landesberger et al. [62], we assign colors to objects according to their positions in a projection reflecting similarities between them. With this approach, color similarity indicates object similarity.

5.3 Relationship Explorer

The Relationship Explorer (Fig. 6a-f) shows the distributions of the data over the spatial, temporal, and topical dimensions and enables the exploration of relationships among these distributions. The main views are the **location view** (Fig. 6a), **time view** (Fig. 6b), and the **topic view** (Fig. 6c) designed according to the uniformity principle (Section 5.1).

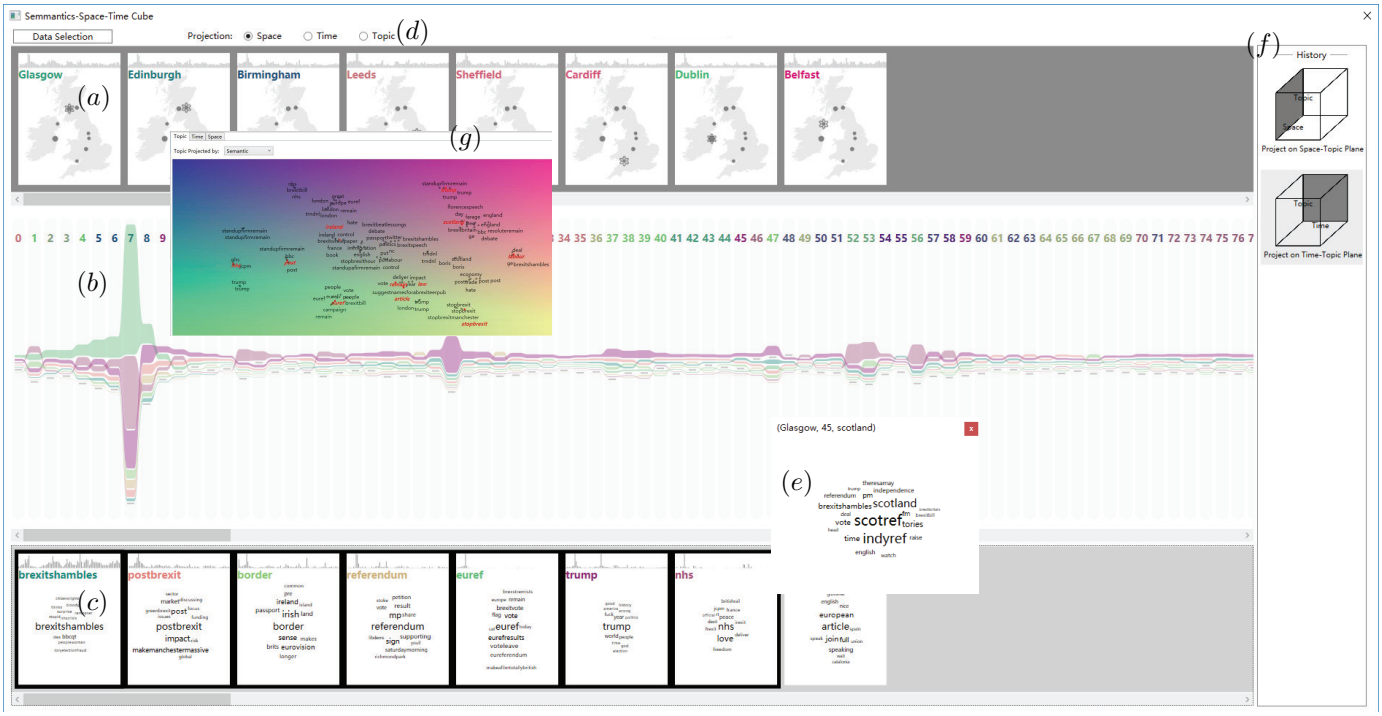


Fig. 6. The visual interface designed for analyzing data involving spatial, temporal, and semantic facets includes two components: Data Selector (g) provides a data overview and enables selection of interesting subsets for exploration. Relationship Explorer (a-f) provides three perspectives on the data based on the facets and supports establishing links among the facets.

The **projection/slice selector** (Fig. 6d) is used to define the current perspective on the data by selecting a projection or slice of the cube (Section 2.3). The history panel (Fig. 6f) represents symbolically what perspective is taken currently and what were taken before. Besides these permanent components, the analyst may create temporary pop-up windows representing the keyword weight vectors for selected cube points (Fig. 6e).

5.3.1 Visual Encodings

The **location view** (Fig. 6a) includes a sequence of location cards showing locations on a map (see Section 5.1-SU). A location card is meant for comparing data at one location, called current location, with data at the other locations. The visual design of a card is shown in Fig. 7a. The card is labeled by the name of the current location colored according to its position in the projection in the Data Selector. A time graph on top of the card shows the temporal variation of the number of documents posted at the current location. The location label is colored according to its position in the projection in the Data Selector (Section 5.2). In the map, the graduated circles represent the counts of documents posted at the respective locations in comparison to the current location using blue for lower values and pink for higher values. The current location is marked with a star symbol.

Similarly to the location view, the **topic view** (Fig. 6c) consists of topic cards, which follows the principle of structure uniformity (Section 5.1-SU). The visual appearance of topic cards (Fig. 7b) reflects the specifics of topics (which are, essentially, vectors of keyword weights) whilst being consistent with the appearance of the location cards, following the principle of visual style uniformity (Section 5.1-VU). The main component of a topic card is a word cloud composed on the topic-related keywords. The font sizes are proportional to the weights of the keywords for the topic. The time graph on top of a topic card shows the popularity variation of the topic. Upon mouse-hovering on a topic card, the keywords that occur in the cards of other topics are highlighted. Simultaneously, these keywords are also highlighted in all cards where they appear. This enables comparing semantic contents of different topics.

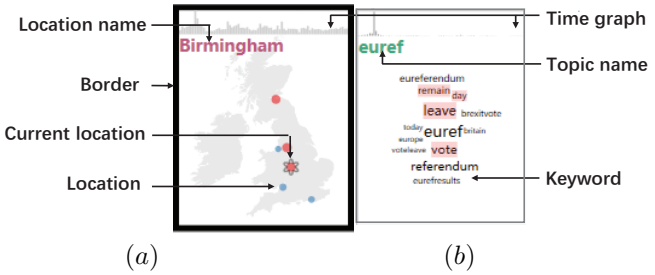


Fig. 7. Visual design of (a) location card and (b) topic card. The card labels are colored according to the positions of the respective location and topic in the projection in the Data Selector.

The **time view** (Fig. 6b) includes narrow cards corresponding to time intervals (Section 5.1-SU) and arranged in a chronological layout reflecting the inherent ordering of the elements of the time dimension. The time view has two modes, compact and extended, see in Fig. 8. In the initial (compact) mode (Fig. 8a), the cards contain multiple vertically arranged bars, which may correspond to topics or locations. The bar colors correspond to the positions of the respective items (i.e., topics or locations) in the projection in the Data Selector (Section 5.2). The bar heights are proportional to the respective popularity scores; the ordering

from top to bottom corresponds to the decreasing order of the scores. One or more items can be selected for viewing the variation of the respective popularity scores in more detail in the extended mode. The bars representing the same selected item in consecutive cards are connected to form a continuous stream, as in Fig. 8b. This representation follows the idea of the ThemeRiver display [18]. The heights of the selected bars extend to use the majority of the vertical space, while the unselected bars are squeezed into a line. The vertical arrangements of all bars do not change to maintain the popularity rankings.

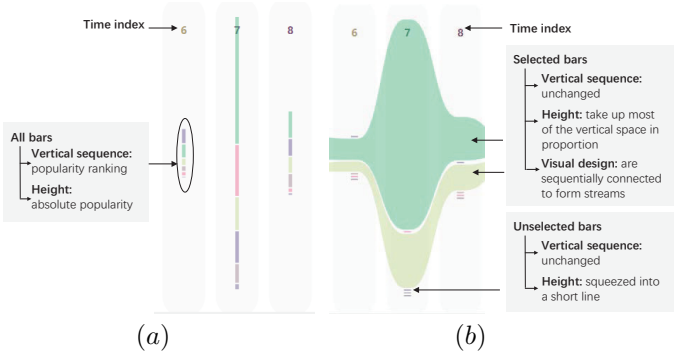


Fig. 8. Two modes of a time view representing 6 topics and 3 time cards. (a) The compact mode. (b) The extended mode, in which the popularity variation of two selected topics is represented by continuous streams.

The **keyword view** is a pop-up window showing the keyword weights for selected cube points (i.e., combinations of location, time, and topic), or aggregated keyword weights for points on cube projections, or even more aggregated weights for elements of a cube dimension (i.e., the values are aggregated over the other two dimensions) (Fig. 9). The information is shown in the form of word cloud with the font sizes proportional to the keyword weights. The analyst can open several such windows, which are linked through interactive highlighting of the same keyword upon mouse-hovering in one of the word clouds.

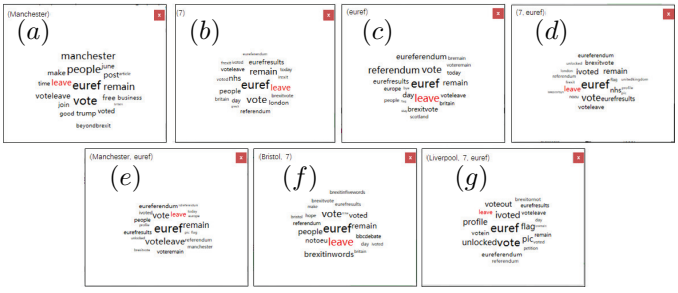


Fig. 9. Keyword views. (a-c): Views showing aggregated keyword weights for elements in three dimensions, a location (a), a time interval (b), and a topic (c). (d-f): Views showing aggregated data for points on 2D cube projections, time-topic (d), location-topic (e), and location-time (f). (g): A view showing detailed data for a cube point, i.e., a combination of location, time, and topic. The same keyword 'leave' is selected for comparing its weights in the different views.

The **history panel**, as in Fig. 6f, symbolically shows the currently chosen perspective of the data, i.e., whether it is a slice or a projection and the kind of slice or projection. It also shows the previous choices in the chronological order from top to bottom and allows the analyst to re-visit the corresponding views through clicking on the icons.

5.3.2 *Interactive Operations for supporting Tasks*

According to the uniformity principle (Section 5.1), we support all task categories **T1** - **T6** defined in Section 2.4 in a uniform way (IU, OU). To conduct a task, the analyst starts with selecting a projection or slice using the Projection Selector. This defines a two-dimensional 'plane', and the analyst's goal is to explore the data distribution over this 'plane'. Please note that the term 'plane' is used metaphorically while the real data structure is more complex; therefore, the visual representation of the selected 'plane' is decomposed into three views. The exploration is performed through selecting objects in one view and observing the corresponding distributions in the other views.

Let us take the task $\Sigma T \rightarrow (S \rightarrow P \rightarrow M)$ (**T1**) as an example. The analyst takes the time projection, i.e., ΣT. Then the analyst selects different locations s in the location view via mouse clicking and observes the corresponding patterns in the the topic view. The interactive operations for the six tasks are summarized in Table 2. The term "pattern" for different tasks has the following meanings:

T1: 1) The topics are sorted according to their popularities at the selected location s; 2) a new topic card showing the popularities of all topics at s is added, as shown in Fig. 10a for s=Dublin and s=Glasgow. The word cloud in the new card consists of the representative words of the topics.

T2: 1) The locations are sorted according to the popularities of the selected topic p; 2) a new location card that shows the spatial distribution of the popularity of p is added, as shown in Fig. 10b for p='scotland' and p='ireland' ('scotland' and 'ireland' are the representative keywords of the selected topics).

T3: For the selected topic p, the time view shows a continuous stream representing the variation of the topic popularity, as in Fig. 10c.

T4: A new topic card showing the popularities of all topics at the selected time interval t is added, as shown in Fig. 10d for t=7 and t=27.

T5: For the selected location s, the time view shows a continuous stream representing the variation of p or ΣP at s, as in Fig. 10e.

T6: A new location card showing the spatial distribution of the popularities of p or ΣP at the selected time interval t is added, as shown in Fig. 10f for $t \in \{54, 55, 56, 57\}$.

As can be seen from the examples in Fig. 10, the analyst can select two or more locations, topics, or time intervals for performing comparison tasks within the categories **T1** - **T6**.

The discussion presented above refers to the patterns of the popularity scores. To explore the patterns of the keyword weight distribution, the analyst generates and compares keyword views for selected locations, topics, or times (Fig. 9).

6 Case Studies

We have tested the utility of our approach on two case studies based on social media posts related to social and natural events.

6.1 Brexit Dataset

The term Brexit (https://en.wikipedia.org/wiki/Brexit) refers to a much disputed plan of the United Kingdom to

TABLE 2
Operations for tasks.

Task	Projection	Spatial List	Time List	Topic List
T1	T	select s	t or ΣT	pattern
T2	T	pattern	t or ΣT	select p
T3	S	s or ΣS	pattern	select p
T4	S	s or ΣS	select t	pattern
T5	P	select s	pattern	p or ΣP
T6	P	pattern	select t	p or ΣP

leave the European Union. We used a dataset containing about 380,000 tweets of more than 70,000 users posted during 78 weeks from May 1, 2016 till October 29, 2017 collected through the Twitter streaming API using a spatial query with the bounding rectangles of the UK and Ireland. We retrieved the tweets containing the keyword "brexit", irrespective of the case, in the texts or hashtags and filtered out the tweets with empty hashtag fields, as we use hashtags for aggregating short texts into longer pseudo-documents (Section 4.2.1). Before applying topic modeling, we converted all texts to lower case. As explained in Section 4.2.2, we ran LDA several times giving different values to the parameter n (topic number): $n = 10, 20, 30, 40$. From the resulting 100 topics, we selected 16 non-redundant topics having clear and relevant meanings (thus, we ignored topics with the most prominent words like 'day', 'great', etc.). The set of locations was constructed by selecting 14 top cities of the UK and Ireland according to the total amounts of posted tweets. The time span was divided into 78 week-long intervals. Finally, a cube with dimensions $14 \times 78 \times 16$ was built.

6.1.1 Verifying Expected Patterns

We check, on the one hand, how our system design supports the task types **T1**-**T6** (Section 2.4), on the other hand, whether it is effective in detecting expectable patterns and known facts concerning people's opinions about Brexit. The exploration we conducted is illustrated in Fig.10.

T1. Fig. 10a: We want to see and compare which topics were popular in Glasgow and Dublin. The reason is that these two cities are always projected close to each other in the data selector regardless of the chosen feature vectors. We take the time projection of the cube; then we click on the cards of Dublin and Glasgow in the location view and obtain the corresponding topic cards in the topic view. These cards show us that the topics "scotland" and "ireland" were the most popular in Glasgow and Dublin, respectively. These are expected patterns, because people are usually more concerned with local affairs. We also observe that the topic "scotland" was quite popular in Dublin. This means that the Twitter users in Ireland might find the discussions concerning Scotland relevant also to Ireland. However, we do not observe a reciprocally high interest to the topic "ireland" in Glasgow.

T2: Fig. 10b: We want to see and compare the spatial distributions of the popularities of the topics "scotland" and "ireland". We select these two topics in the topic view and obtain two new cards in the location view showing the spatial distributions of the topic popularities represented by the sizes of the circles. Not surprisingly, we observe that, among all locations, the topic "scotland" was the most popular in Glasgow and Edinburgh while "ireland" was popular in Dublin. People in the other cities were not very much interested in any of these two topics.

T3. Fig. 10c: We want to check whether the times of peaks in topic popularity corresponds to the times of events these topics are related to. This task does not involve the spatial dimension; so, we take the space projection of the cube. We select the topics "euref" and "trump", which refer to the British Brexit referendum and the current US president. In the time view, we observe the popularity variations of these two topics. As could be expected, the highest popularity of "euref" was attained in the 7th week (19-26 June, 2016), when the Brexit referendum was held. The popularity gradually increased before that week and gradually decreased in the following weeks; in the remaining times, it was quite low. The topic "trump" had its highest popularity in the 27th week (6 June, 2016 to 13 June, 2016), which corresponds to the beginning of the US presidential election. Interestingly, this topic was also quite popular at the time of the Brexit referendum and in the following weeks.

T4. Fig. 10d: In the time view, we select the 7th and 27th weeks, in which "euref" and "trump", respectively, had their highest rankings. Two cards showing the popularities of all topics in these two weeks are added to the topic list. The words representing the topics "euref" and "trump" have the largest font sizes corresponding to the highest popularities of these topics in the respective weeks, which is consistent with the previous observations made in the time view. In addition, we observe that, apart from the main topic "euref", the topics "scotland", "remain", "trump", and "stopbrexit" were also quite popular in week 7, whereas week 27 was strongly dominated by the topic "trump".

T5. Fig. 10e: We want to look at the overall Twitter activities regardless of specific topics; therefore, we take the topic projection of the cube. We select five large cities from different parts of the studied territory and use the time view to explore the temporal patterns of the aggregated popularity, which reflects the general activity of the discussions in Twitter. We see a peak of activity in the 7th week (June 20-26, 2016), when the Brexit referendum was held. It shows that the overall number of tweets posted during the referendum period was higher than in the other weeks, thus resulting in a high aggregated popularity. We can also notice that the temporal trends were very similar in all selected cities except Dublin, where the Twitter activity sharply dropped down in week 8 but then increased in weeks 14-15 (August 8-21, 2016), when the activities in the other cities were quite low. The increase of the discussions in Dublin can be related to events in the Northern Ireland, where on August 10 the first and deputy first ministers sent an open letter to the UK prime minister Theresa May concerning Brexit impact on the Northern Ireland. Evidently, people in Ireland, in particular, Dublin, are concerned about the affairs in the Northern Ireland. These observations show that social media activities increase in response to important social events [3].

T6. Fig. 10f: To check the slice-based functionality, we take the cube slice corresponding to the topic "ge", which refers to the British general election. The election was announced on April 18 (week 50) and held on June 8 (week 57). The time view shows us the popularity dynamics of this topic in multiple selected cities. After a peak in the week of

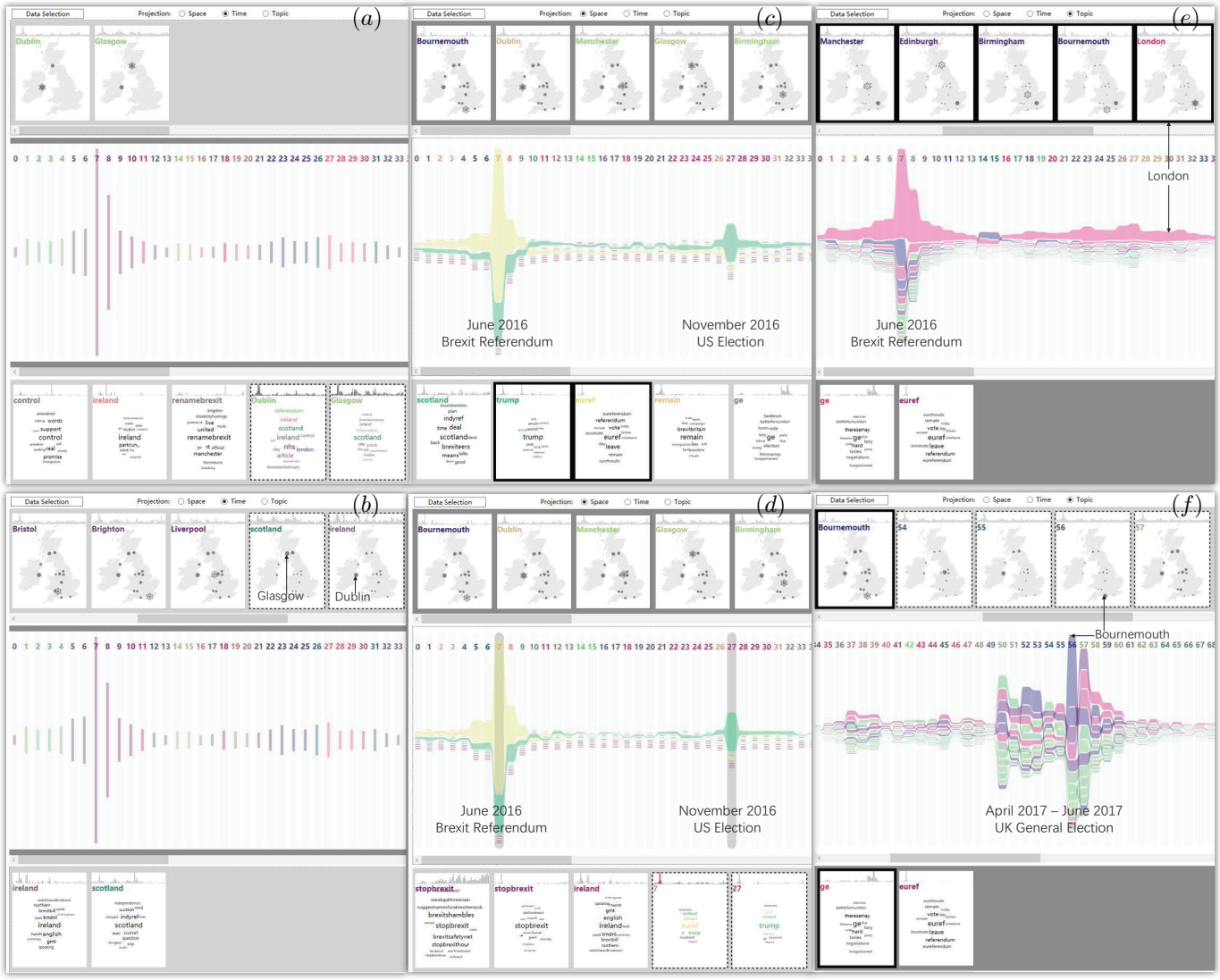


Fig. 10. Demonstration of fulfilling six types of tasks. (a)T1: Take the time projection and select locations to observe patterns in the topic view. (b) T2: Take the time projection and select topics to observe patterns in the location view. (c) T3: Take the space projection and select topics to observe their streams in the time view. (d) T4: Take the space projection and select time intervals to observe the topic patterns in the topic view. (e) T5: Take the topic projection and select locations to observe their streams in the time view. (f) T6: Take a topic slice and select time intervals to observe the spatial distributions in the location view.

election announcement, the activity decreased but remained high and then raised to its highest values in weeks 56-57. We select the weeks 54-57 in the topic view. Four cards showing the spatial distributions of the popularity of "ge" in these weeks appear in the location list. Since the topic had the highest popularity in London in all four weeks, the spatial distribution maps exhibit high domination of London while the circles representing the values in the other cities are too small for seeing differences. Therefore, we exclude London from the selection and look at the spatial patterns formed by the remaining cities. As could be expected, there was a general increase of activities in all cities in the week of the election (week 57). In the preceding weeks, some cities had relatively higher activities than others; however, the activities in weeks 54 and 55 were not very high in absolute values, as can be seen in the time view. Week 56 differs from the others by having a high peak of activity in Bournemouth, on the south of England. We create a corresponding keyword view, which points at an intensive media campaign for supporting the labor party. Bournemouth is the location of the headquarters of the company JP Morgan known for its strong anti-Brexit position. In the week before the general election, they announced that the conservative party losing the general election would be beneficial for the UK finances, which may explain the high number of the pro-labor tweets.

This test has confirmed that, first, all task types are supported, second, they are performed in uniform ways by applying the same interactive techniques in the different views and, third, the techniques have the power to reveal meaningful patterns.

6.1.2 *Studying Temporal Evolution of a Topic*

We select the space projection and the topic "trump" that refers to the current US president, to analyze its temporal evolution. Due to the strong association between US and

UK, a large number of tweets in this dataset discuss the effects of US social events on the Brexit.

There are two peaks in the stream of the topic "trump". The first peak was in the period of the Brexit referendum. We analyzed the corresponding keyword lists and did not find any special keyword (apart from "trump") except the keywords related to the referendum. Hence, the first peak was caused by the popularity accumulation from the large number of tweets posted in that week. The second peak was at the beginning of the US election (6 Nov, 2016 to 12 Nov, 2016). In the keyword view, several keywords related to the election had higher prominences, such as "uselection", "electionnight", and "electionday". Like in the previous examples, this confirms that social media activities are highly related to to important social events. In the 33th week (18 Dec, 2016 to 24 Dec, 2016), the keyword "win" appeared, and the corresponding event was that Trump won the election. We found the keyword "muslimban" in the 39th week (29 Jan, 2017 to 4 Feb, 2017), one day after Trump signed the executive order 13769, which prohibits citizens of seven countries in the Middle East from entering the United States within the next 90 days. The higher keyword prominence indicated active discussions concerning the order in the social media. In the 47th week (26 Mar, 2017 to 1 Apr, 2017), the keyword "article" appeared, which was consistent with the news that the prime minister triggered Article 50 of the Treaty on EU. In the 51th week (23 Apr, 2017 to 39 Apr, 2017) keywords about the election, such as "ge" and "generalelection", appeared again due to the announcement of the British general election. In the 58th week (11 Jun, 2017 to 17 Jun, 2017), we found many names of countries and politicians, such as "russia", "syria", "israel", "usa", "putin", etc., which may be related to the Trump's press conference on June 14, 2017.

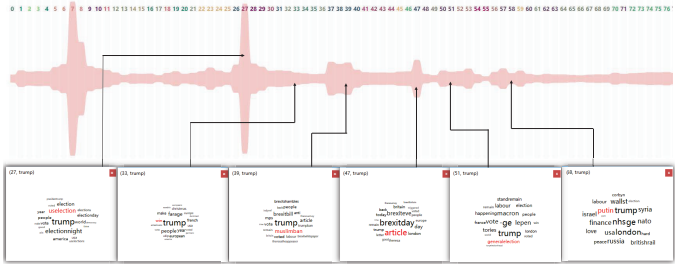


Fig. 11. Temporal evolution of topic "trump".

6.1.3 Exploring Competition of Topics

We compare the temporal trends of two topics with similar meaning, "remain" and "stopbrexit". We find (Fig. 12a) that the popularity of the topic "remain" (the green stream) had been higher than the topic "stopbrexit" (the pink stream) during the first half of the period. Afterwards, the popularity of the topic "stopbrexit" gradually increased and finally exceeded the first one. The topic "stopbrexit", evidently, refers to the development of the Brexit process whereas the topic "remain" refers more to making the Brexit decision. For seeing more details, we create keyword views for different weeks; the keyword views for the weeks 7 and 74 are shown in Fig. 12a. In week 7, many keywords related to the referendum and voting had higher prominence within the topic "remain", indicating a willingness to vote for Britain to stay in EU. In week 74, the topic "stopbrexit" contained many keywords related to Manchester, indicating some local events in Manchester. Therefore we selected the location slice corresponding to Manchester and observed the corresponding popularity variations of the two topics (Fig. 12b). We found the same temporal trends as for all locations in general, except for a more significant popularity difference in week 74 (Oct 1 to Oct 7, 2017). By retrieving the related information from the Internet, we learned that an organization named *stopbrexit* https://www.stopbrexitmarch.com/ organized a march on October 1 2017 in Manchester, thus causing the popularity burst. We also found the keyword "march" in the keyword distribution. The increase of the popularity of "stopbrexit" in week 70, which can be seen in the time view in Fig. 12a, is related to a similar anti-Brexit match that happened on September 9 in London. Hence, by investigating the temporal variation patterns, we can find references to important local events.

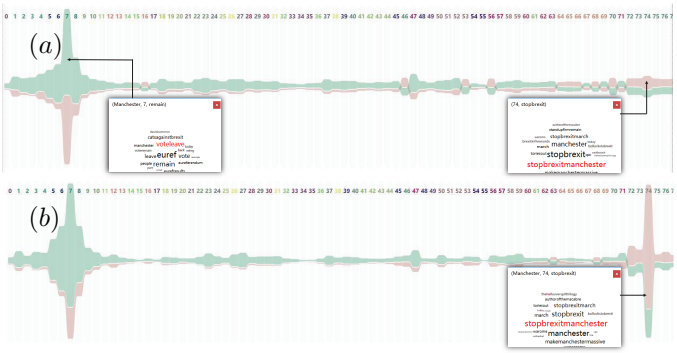


Fig. 12. Different popularity trends of "remain" (green) and "stopbrexit" (pink): a) the whole territory, (b) slice for "Manchester".

6.1.4 Interesting Findings regarding the Spatial Distribution

We take the time projection and generate location cards showing the spatial distributions of the popularities of different topics, as in Fig. 13a, b. We also generate topic cards showing the topic popularities distribution in different selected cities, as in Fig. 13c, d. In this way, we find answers to interesting questions regarding $space - topic$ relations. A few examples are presented below.

Which city has the highest concerns about the situation after the Brexit?. We look at the spatial distribution of the popularity of the topic "postbrexit" and see that the highest activity was in Dublin (Fig. 13a). We create a corresponding keyword view (Fig. 13e), where we see such keywords as "impact", "market", "discussing", etc. We learn from literature that currently there is no actual border between North Ireland and Ireland, but this situation may change after the Brexit, which will seriously affect the trade between Ireland and Britain. This explains the high concern about the topic "postbrexit" in Dublin.

Which area has the highest opposition to the Brexit? We look at the spatial distribution of the popularity of the topic "brexitshambles" (Brexit Shambles), which reflects a dissatisfaction with the chaos caused by the Brexit. The topic was the most popular in Ireland and Scotland; particularly,

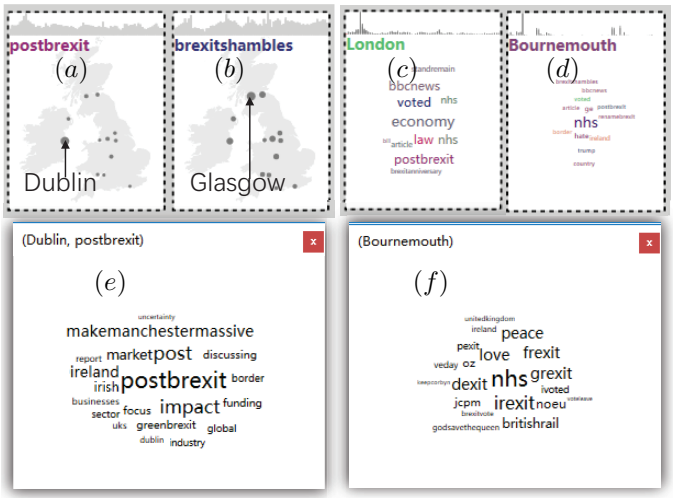


Fig. 13. Interesting findings in respect to the spatial distributions of the topic popularities.

Glasgow had the highest value (Fig. 13b). The corresponding keyword view contains the keywords "indyref" and "snp" , which respectively refer to the Scotland independence referendum and the party leading the independence campaign.

Which topics were the most interesting to people in London? The topic card presented in Fig. 13c shows us that the topic "economy" received the highest attention in London. This can be easily explained, since the Brexit has a high impact on economy. Other topics, such as "bbcnews", "postbrexit", "law", etc., also received much attention. As London is the capital of the UK and also the largest city, the London's distribution of the topic popularities may be indicative of the interests and priorities of the whole British society.

Which city was the most different from the others regarding the topic popularities? By comparing the topic cards of different cities, we find that, indeed, the respective topic popularities were quite similar to those in London. A single exception is Bournemouth, where the most popular topic was NHS (this is an abbreviation of the National Health System), and it strongly dominated all other topics (Fig. 13d). In the corresponding keyword view (Fig. 13f), we observe a very high prominence of the keyword 'nhs'. The NHS topic corresponds to the fears that the Brexit will have a very negative impact on the British public health system. This topic appears also in the cards of the other cities, but its relative popularity is much lower than in Bournemouth. The specificity of Bournemouth can be related to the fact that it is a place of a major hospital belonging to the NHS Foundation. The popularity of the NHS topic may reflect concerns of the hospital employees and/or patients.

The examples from the Brexit case study demonstrate the effectiveness of our approach for studying the relationships between text semantics, space, and time.

6.2 Storm Doris Dataset

This case study demonstrates analysis at a small temporal scale. Storm Doris hit Ireland and UK on February 23-24, 2017. During these two days, people actively shared their storm-related experiences in social media. We retrieved the messages containing such keywords as "storm", "rain", "wind", and "snow" in the texts or hashtags. In total, 30,000 tweets of about 25,000 users were collected. Most of these tweets were posted in three cities, Dublin, Cardiff and Glasgow; so, we selected these cities for the analysis (Fig. 14a). The time period was divided into 48 intervals of one hour length. We utilized city-time aggregation for topic extraction, i.e. the tweets posted within one hour in a city were merged to form a longer pseudo-document as the input to the LDA algorithm (see Section 4.2.1). Compared to the Brexit dataset, this dataset contains fewer topics, since most tweets either mention the meteorological conditions or refer to emergency notices and latest news, such as flight delays and accidents leading to casualties. Therefore we selected 5 representative topics. Based on these selections, a cube of $3 \times 48 \times 5$ was generated for the exploration.

Our goal is to trace the storm in the social media, similarly to earlier studies related to other natural events [22], [23]. Specifically, we would like to observe whether the times of the popularity bursts of the storm-related topics in different cities are consistent with the trajectory of the storm. We want to check the hypothesis that people tend to tweet immediately about their current experiences. We know from news reports that heavy snow and rain occurred only in North Britain whereas strong winds affected a larger territory. Therefore, we take the slice of the topic "windy" (Fig. 14b), which refers to the windy weather, to explore its spatio-temporal dynamics. We find that the popularity of the topic first peaked in Dublin at 7am on the 23th, then a burst successively occurred in Cardiff in the following hour, and finally a burst occurred in Glasgow at 1pm, as can be seen in Fig. 14c. We marked the cities on a weather forecast map, as in Fig. 14d. By sequentially connecting the cities, we found that the sequence of the occurrences of the topic popularity bursts is consistent with the wind directions and storm movement. A video analyzing the spatial impacts of Storm Doris can be found in https://www.metoffice.gov.uk/barometer/uk-storm-centre/storm-doris/. We learned from it that the wind was from the northwest to the southeast, and Scotland was the region where strong wind appeared the latest. This finding proves our hypothesis.

7 Expert Feedback

For an additional evaluation of the potential usefulness of our approach, we established a remote contact with an expert from the Free University of Berlin, whose research interest is political analysis, in particular, studying the relationships between politics and social media. Our goal was to test whether our system can be helpful to the expert in her professional research. We used a communication platform supporting screen sharing to demonstrate the analysis of the Brexit data and explain the essentials of our approach. In the course of the demonstration, we answered emerging expert's questions, which involved creating views that showed requested information or interacting with already available views. We asked the expert to make comments during the demonstration and to provide an overall judgment after it.

The overall feedback was positive: *"It is quite easy to understand and the uniform organization of topic, space and time*

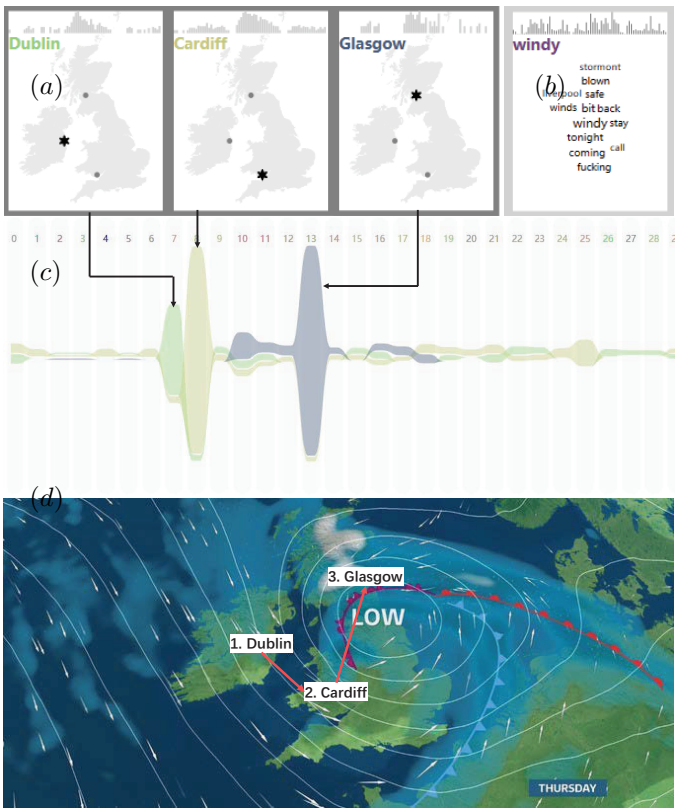


Fig. 14. Spatiotemporal dynamic of the topic "windy" extracted from Storm Doris dataset. (a) Three selected cities. (b) Topic "windy". (c)Streams show the popularity trend of the three cities with respect to the topic "windy". (d) Weather forecast map that shows the wind movement trends on 23th.

information is impressive. In our research, we also addressed part of the analysis tasks, such as analyzing evolutionary pattern of topics (topic and time), the opinion distribution over different locations (space and topic). But we didn't have the tool to organize in such way. The cube is a good summarize metaphor and extends the capability of political analysis. It can be applied in multiple fields in our domain."

As to the visual design and interactive operations, the expert said the the Relationship Explorer is not as easy-to-understand as the Data Selector. However, she admitted that the the design uniformity helped a lot in understanding and learning the tool. She said: *"The interface is very beautiful and user-friendly. I can easily understand the Data Selector. The Relationship Explorer is more complex for me, since I should always keep the concept of the cube in my mind. The learning curve of the Relationship Explorer is steep. When the user understand the task scheme, the use of the Relationship Explore will become easier, since all the tasks follow the same scheme."*

She also made some detailed comments, both positive and negative: *"The keyword view is very helpful, without it the popularity variation is difficult to explain. The two measures are well selected. I like the complete task definition, and to integrate them in a tool is very convenient to use. We need to pay attention that the number of specific mentioned keywords may not really reflect the public opinions. For example, if I mention Stopbrexit, I may tweet for being against Stopbrexit. I hope in the future you can provide further supports on the detailed semantic analysis. "*

The expert expressed a desire to use our approach in her political science research. She said: *"We always met such problems: For political events, such as elections, we especially care about the people's caring topics in indecisive states in specific time periods. I believe the proposed visual analytics can help a lot."*

Hence, the expert's feedback was very encouraging while pointing at some aspects requiring improvement.

8 Discussion

In this paper, we describe an approach to exploring data containing texts associated with spatial locations and times. While we presented our approach by example of social media data, it can be applied to data with similar structure existing in other domains, for example, reports about incidents, crimes, or technical failures of vehicles [45]. A specific feature of our approach is the use of the data cube metaphor for organizing data and in the design of the interactive visual interface. This metaphor implies uniform treatment of different data components. On this basis, we developed the idea of uniformity in application to the visualization and interaction design with the purpose to mitigate the complexity of the data and the exploratory tasks. The cube metaphor also allowed us to define systematically the space of possible tasks in data exploration. In the following, we discuss different aspects of our approach.

Topic extraction. We utilize topic modeling, specifically, LDA, to represent unstructured texts in a structured way. We aggregate texts with similar semantics into longer pseudo-documents to deal with the problem of term scarcity in short texts. Another problem is setting the number of topics to extract, which may be difficult to estimate. As a viable approach, we propose to run LDA several times with different parameter settings and present the resulting topics in a projection view exhibiting their similarities, allowing the analyst to select a non-redundant set of clearly interpretable topics relevant to the analysis goals.

Measure calculation. In our research, we used two heuristically defined measures, namely, topic popularity score and keyword weight vector. It is also possible to define and use other measures, depending on the specifics of the application domain and the goals of the analysis. Thus, the feedback of the expert in political analysis (Section 7) indicated the usefulness of a measure reflecting people's attitudes towards the topics (which would not be needed in analyzing crime or incident reports). As mentioned in Section 2.3, such a measure could be defined based on outcomes of text sentiment analysis. Inclusion of additional measures in the cube, in principle, does not affect the overall approach, but the visual interface may need to be extended for presenting these measures.

Task completeness. We have defined the complete system of analysis tasks that may refer to data organized in an SSTC. However, this includes only tasks applicable to aggregated data but not tasks that require dealing with detailed text-time-space data, such as the task of cluster detection addressed in earlier works [22], [23], [32].

Information loss. Since cube construction involves discretization (binning) and aggregation of original data, it inevitably entails information loss. On the other hand, aggregation allows the analyst to disregard minor details and

supports abstractive grasp of essential features. As with any kind of aggregation, the amount of information that is lost depends on the granularity, i.e., the sizes of the bins that are used. Suitable bin sizes are chosen depending on the scope and variation of the data, the scale of the phenomenon that is studied, and the required scale of analysis. Thus, in our two case studies, we used weekly time intervals for studying a long-term phenomenon and hourly intervals for a short-living phenomenon.

Visualization effectiveness. We have designed a visual interactive interface that supports all previously defined task types, as was proved in the case studies. Moreover, we developed and consistently applied the principle of uniformity, so that different types of tasks can be performed in a uniform way. The uniformity as a means of decreasing system complexity and reducing the learning effort received positive comments in the expert evaluation. However, the evaluation revealed some issues concerning the way in which the uniformity principle had been implemented. These issues are discussed in the following paragraph.

Simplicity and ease of use. Our underlying idea in designing the user interface was explicit use of the cube metaphor and the concepts of cube dimension, slice, and projection. The metaphor was even emphasized by the icons appearing in the history panel. We expected that this idea will promote understanding of the system and make it easier to use. However, the expert evaluation showed that thinking of data in terms of the cube metaphor may not be as obvious and easy to a user as we expected. Hence, while the uniformity principle is useful and should be preserved, it is appropriate to find a different approach to implementing it in the UI design. Currently, we consider an idea of organizing the UI based on the six task types, i.e., the analyst selects the type of task he/she is going to perform from a task list. For this purpose, the tasks need to be given short but well-understandable and unambiguous titles. It would be good to design such titles in communication with several analysts, preferably from different domains. This would include testing the possibility to title the tasks in general, domain-independent terms.

Performance. We divided the workflow into two phases. The first includes topic extraction and data cube construction, which are computationally intensive and time consuming operations. The second is interactive data exploration. The separation of the data preparation from the exploration minimizes the CPU and memory usage. Moreover, our system establishes various references among data items using hash structures for accessing the cube components. The initialization of the cube (i.e., preparation to the interactive exploration) takes several seconds and may become longer as the numbers of elements increase. The initialization step, however, improves the system performance during the exploration phase and enables real time responses to interactive operations.

Scalability. While the amount of original data may be huge, it is substantially reduced in constructing the cube due to aggregation. Generally, a cube is a scalable structure, which makes it widely used. Still, when cube dimensions consist of large numbers of elements, the cube may be too "heavy" for keeping in the main memory and interactive exploration. This problem can be mitigated by applying the Data Selector for selecting sub-cubes for further exploration in the Relationship Explorer. As a data overview, the Data Selector can accommodate a large number of objects, from which analysts can select subsets of interest. What concerns the visualization scalability, the capacity of the displays to show multiple components (such as cards) simultaneously is obviously limited. While all views include scroll bars, it may be inconvenient and distracting from the analysis to scroll repeatedly for finding items of interest. This problem is alleviated by meaningful ordering of the items within views and by providing the possibilities for removing and adding items according to the current focus.

Practical usefulness. SSTC has practical utility first of all as a data structure suitable for representing the distribution of text semantics over space and time and thereby enabling analysis tasks that have not been supported before. It is worth noting that this data structure allows application of the previously proposed approaches to performance optimization, such as NanoCubes [40]. The SSTC is also useful as a conceptual model, as it provides a ground for a systematic and comprehensive definition of the space of possible analysis tasks and, simultaneously, determines the database operations that are necessary for supporting these tasks. However, as discussed above, SSTC may be less suitable as a metaphor for organizing the visual user interface. This conclusion from our study may be useful to other researchers and visualization designers.

9 Conclusion

This paper has presented a comprehensive approach to exploring data with spatial, temporal and textual components. The approach can be summarized as follows.

1) A data cube structure is used for organizing such data. We describe how the original data are transformed and organized in a semantics-space-time cube.
2) Topic modeling is used for converting unstructured texts into a structured representation. We propose a viable approach (based on using an LDA ensemble) to compensating for parameter impact and extracting a meaningful and useful set of topics.
3) Based on the cube structure, the space of possible exploratory analysis tasks for topic-space-time data is systematically and comprehensively defined.
4) The principle of design uniformity aims at supporting all task types in a uniform way to moderate the intrinsic complexity.
5) We demonstrate a possible implementation of the uniformity principle in a visualization system design.

We tested the system implementing our approach in two case studies to confirm that all task types are supported and the uniformity principle works appropriately. We also undertook expert evaluation for testing the usability aspects of the approach and received valuable feedback, which will be taken into account in the further work. In particular, since users may have difficulties with explicit use of the cube metaphor, we shall try to create a different variant of the UI design incorporating the uniformity principle, since this principle was acknowledged as useful and important.

ACKNOWLEDGMENTS

The work is partially supported by National NSFC projects (Grant numbers 61602340 and 61572348), by National High-tech R&D Program (863 Grant number 215AA020506), by Fraunhofer Cluster of Excellence on "Cognitive Internet Technologies", by EU in projects DiSIEM and SoBigData and by DFG (German Research Foundation) in priority research program SPP 1894 "Volunteered Geographic Information: Interpretation, Visualization and Social Computing".

REFERENCES

[1] J. Gray, A. Bosworth, A. Lyaman, and H. Pirahesh, "Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals," in *Proceedings of the Twelfth International Conference on Data Engineering*, Feb 1996, pp. 152–159.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[3] J. Li, S. Chen, G. Andrienko, and N. Andrienko, "Visual exploration of spatial and temporal variations of tweet topic popularity," in *Proceedings of EuroVA' 18*, 2018.

[4] N. Andrienko and G. Andrienko, *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Berlin, Heidelberg: Springer-Verlag, 2005.

[5] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, January 2008.

[6] G. Salton and M. J. McGill, "Introduction to modern information retrieval," 1986.

[7] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999.

[8] J. Choo, C. Lee, C. K. Reddy, and H. Park, "UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 1992–2001, 2013.

[9] H. Lee, J. Kihm, J. Choo, J. T. Stasko, and H. Park, "ivisclustering: An interactive visual document clustering via topic modeling." *Computer Graphics Forum*, vol. 31, no. 3, pp. 1155–1164, 2012.

[10] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian, "Interactive, topic-based visual text summarization and analysis," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ser. CIKM '09. ACM, 2009, pp. 543–552.

[11] S. Chen, L. Lin, and X. Yuan, "Social Media Visual Analytics," *Computer Graphics Forum*, vol. 36, no. 3, pp. 563–587, 2017.

[12] X. Wang, W. Dou, Z. Ma, J. Villalobos, Y. Chen, T. Kraft, and W. Ribarsky, "*I-SI*: Scalable architecture for analyzing latent topical-level information from social media data," *Computer Graphics Forum*, vol. 31, no. 3, pp. 1275–1284, 2012.

[13] P. Xu, Y. Wu, E. Wei, T.-Q. Peng, S. Liu, J. J. Zhu, and H. Qu, "Visual analysis of topic competition on social media," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2012–2021, 2013.

[14] G. Sun, Y. Wu, S. Liu, T.-Q. Peng, J. J. Zhu, and R. Liang, "Evoriver: Visual analysis of topic coopetition on social media," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1753–1762, 2014.

[15] X. Wang, S. Liu, Y. Chen, T.-Q. Peng, J. Su, J. Yang, and B. Guo, "How ideas flow across multiple social groups," in *Proceedings of IEEE VAST' 16*, 2016, pp. 51–60.

[16] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky, "Hierarchicaltopics: Visually exploring large text collections using topic hierarchies," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2002–2011, 2013.

[17] W. Cui, S. Liu, Z. Wu, and H. Wei, "How hierarchical topics evolve in large text corpora," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 2281–2290, 2014.

[18] S. Havre, B. Hetzler, and L. Nowell, "Themeriver: Visualizing theme changes over time," in *Proceedings of IV*, 2000, pp. 115–123.

[19] Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu, "Opinionflow: Visual analysis of opinion diffusion on social media," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1763–1772, 2014.

[20] J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin, and C. Collins, "# fluxflow: Visual analysis of anomalous information spreading on social media," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1773–1782, 2014.

[21] S. Liu, J. Yin, X. Wang, W. Cui, K. Cao, and J. Pei, "Online visual analytics of text streams," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 11, pp. 2451–2466, 2016.

[22] N. Andrienko, G. Andrienko, G. Fuchs, S. Rinzivillo, and H.-D. Betz, "Detection, tracking, and visualization of spatial event clusters for real time monitoring," in *Proceedings of IEEE DSAA' 15*, 2015, pp. 1–10.

[23] G. Fuchs, N. Andrienko, G. Andrienko, S. Bothe, and H. Stange, "Tracing the german centennial flood in the stream of tweets: first lessons learned," in *Proc. 2nd ACM SIGSPATIAL workshop on Crowdsourced and Volunteered Geogr. Information*, 2013, pp. 31–38.

[24] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th intl conference on World Wide Web*, 2010, pp. 851–860.

[25] W. Dou, X. Wang, W. Ribarsky, and M. Zhou, "Event detection in social media data," pp. 971–980, 01 2012.

[26] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller, "Twitinfo: aggregating and visualizing microblogs for event exploration," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2011, pp. 227–236.

[27] X. Zhou and C. Xu, "Tracing the spatial-temporal evolution of events based on social media data," *ISPRS International Journal of Geo-Information*, vol. 6, no. 3, p. 88, 2017.

[28] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl, "Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition," in *Proceedings of IEEE VAST'02*, 2012, pp. 143–152.

[29] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou, "Leadline: Interactive visual analysis of text data through event identification and exploration," in *Proceedings of IEEE VAST' 12*, 2012, pp. 93–102.

[30] I. Cho, R. Wesslen, S. Volkova, W. Ribarsky, and W. Dou, "Crystalball: A visual analytic system for future event discovery and analysis from social media data," in *Proc. of IEEE VAST'17*, 2017.

[31] T. Kraft, D. X. Wang, J. Delawder, W. Dou, Y. Li, and W. Ribarsky, "Less after-the-fact: Investigative visual analysis of events from streaming twitter," in *Proc. IEEE LDAV'13*, 2013, pp. 95–103.

[32] H. Bosch, D. Thom, M. Wörner, S. Koch, E. Püttmann, D. Jäckle, and T. Ertl, "ScatterBlogs: Geo-spatial document analysis," in *Proceedings of IEEE VAST' 11*, 2011, pp. 309–310.

[33] H. Bosch, D. Thom, F. Heimerl, E. Püttmann, S. Koch, R. Krüger, M. Wörner, and T. Ertl, "Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2022–2031, 2013.

[34] J. Chae, D. Thom, Y. Jang, S. Kim, T. Ertl, and D. S. Ebert, "Public behavior response analysis in disaster events utilizing visual analytics of microblog data," *Computers & Graphics*, vol. 38, pp. 51–60, 2014.

[35] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, and H. Qu, "Whisper: Tracing the spatiotemporal process of information diffusion in real time," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2649–2658, 2012.

[36] G. Andrienko, N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, and D. Thom, "Thematic patterns in georeferenced tweets through space-time visual analytics," *Computing in Science & Engineering*, vol. 15, no. 3, pp. 72–82, 2013.

[37] G. Cao, S. Wang, M. Hwang, A. Padmanabhan, Z. Zhang, and K. Soltani, "A scalable framework for spatiotemporal analysis of location-based social media data," *Computers, Environment and Urban Systems*, vol. 51, pp. 70–82, 2015.

[38] S. Shekhar, C. T. Lu, R. Liu, and C. Zhou, "Cubeview: a system for traffic data visualization," in *Proceedings of the 5th IEEE Intelligent Transportation Systems.*, 2002, pp. 674–678.

[39] Y. Tian, R. A. Hankins, and J. M. Patel, "Efficient aggregation for graph summarization," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 567–580.

[40] L. Lins, J. T. Klosowski, and C. Scheidegger, "Nanocubes for real-time exploration of spatiotemporal datasets," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2456–2465, 2013.

[41] C. A. Pahins, S. A. Stephens, C. Scheidegger, and J. L. Comba, "Hashedcubes: Simple, low memory, real-time visual exploration of big data," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 671–680, 2017.

[42] Z. Wang, N. Ferreira, Y. Wei, A. S. Bhaskar, and C. Scheidegger, "Gaussian cubes: Real-time modeling for visual exploration of large multidimensional datasets," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 681–690, 2017.

[43] F. Miranda, M. Lage, H. Doraiswamy, C. Mydlarz, J. Salamon, Y. Lockerman, J. Freire, and C. Silva, "Time lattice: A data structure for the interactive visual analysis of large time series," in *Proceedings of EuroVis' 18*, vol. 37, no. 3, 2018, pp. 13–22.
[44] C. X. Lin, B. Ding, J. Han, F. Zhu, and B. Zhao, "Text cube: Computing ir measures for multidimensional text database analysis," in *Proceedings of ICDM'08*, 2008, pp. 905–910.
[45] D. Zhang, C. Zhai, J. Han, A. Srivastava, and N. Oza, "Topic modeling for OLAP on multidimensional text databases: Topic cube and its applications," *Statistical Analysis and Data Mining*, vol. 2, no. 5-6, pp. 378–395, December 2009.
[46] T. Hägerstraand, "What about people in regional science?" *Papers in regional science*, vol. 24, no. 1, pp. 7–24, 1970.
[47] P. Gatalsky, N. Andrienko, and G. Andrienko, "Interactive analysis of event data using space-time cube," in *Proceedings of IV' 04*, 2004, pp. 145–152.
[48] T. Kapler and W. Wright, "Geotime information visualization," in *IEEE Symposium on Information Visualization 2004(INFVIS)*, vol. 00, 2007, pp. 25–32.
[49] N. Andrienko and G. Andrienko, "Visual analytics of movement: an overview of methods, tools, and procedures," *Information Visualization*, vol. 12, no. 1, pp. 3–24, 2013.
[50] B. Bach, P. Dragicevic, D. Archambault, C. Hurter, and S. Carpendale, "A review of temporal data visualizations based on space-time cube operations," in *Proceedings of Eurographics conference on visualization*, 2014.
[51] F. Amini, S. Rufiange, Z. Hossain, Q. Ventura, P. Irani, and M. J. McGuffin, "The impact of interactivity on comprehending 2d and 3d visualizations of movement data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, pp. 122–135, 2015.
[52] S. Dübel, M. Röhlig, H. Schumann, and M. Trapp, "2d and 3d presentation of spatial data: A systematic review," in *2014 IEEE VIS International Workshop on 3DVis (3DVis)*, Nov 2014, pp. 11–18.
[53] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel, *Visual Analytics of Movement*. Springer, 2013.
[54] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 1445–1456.
[55] X. Quan, C. Kit, Y. Ge, and S. J. Pan, "Short and sparse text topic modeling via self-aggregation." in *Proceedings of IJCAI' 15*, 2015, pp. 2270–2276.
[56] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of 3rd ACM conference on Web search and data mining*, 2010, pp. 261–270.
[57] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the first workshop on social media analytics*, 2010, pp. 80–88.
[58] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
[59] J. Kruskal and M. Wish, *Multidimensional Scaling*, ser. 07. Sage Publications, Inc, 1978, no. 11.
[60] N. Andrienko, G. Andrienko, G. Fuchs, and P. Jankowski, "Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces," *Information Visualization*, vol. 15, no. 2, pp. 117–153, 2016.
[61] S. Haroz and D. Whitney, "How capacity limits of attention influence information visualization effectiveness," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2402–2410, 2012.
[62] T. Von Landesberger, S. Bremm, N. Andrienko, G. Andrienko, and M. Tekusova, "Visual analytics methods for categoric spatio-temporal data," in *Proceedings of IEEE VAST'12*, 2012, pp. 183–192.

Jie Li is an assistant professor at College of Intelligence and Computing of Tianjin University. He got his PhD from Tianjin University. After that he visited at Fraunhofer IAIS as a PostDoc researchers. His current research interests include visual analytics of social media, public security and environmental science. He published several papers in IEEE TVCG, IEEE VIS, Journal of Computer, etc. More information can be found at http://geova.cn/jieli/.

Siming Chen is a research scientist at Fraunhofer Institute IAIS and a PostDoc researcher of University of Bonn in Germany. He got his PhD from Peking University. His research interests include visual analytics of social media, cyber security and spatial temporal data. He published several papers in IEEE VIS, IEEE TVCG, EuroVis, etc. More information can be found at http://simingchen.me.

Wei Chen is a professor in State Key Lab of CAD&CG at Zhejiang University. His current research interests include visualization and visual analytics. He has published more than 40 IEEE/ACM Transactions and IEEE VIS papers. He served in many leading conferences and journals, like IEEE PacificVIS, ChinaVIS steering committee, IEEE LDAV and ACM SIGGRAPH Asia VisSym. He is also the associate EIC of JVLC, associate editor of IEEE CG&A and JOV.

Gennady Andrienko is a lead scientist responsible for visual analytics research at Fraunhofer Institute for Intelligent Analysis and Information Systems and part-time professor at City University London. Gennady Andrienko was a paper chair of *IEEE VAST* conference (2015–2016) and associate editor of *IEEE Transactions on Visualization and Computer Graphics* (2012–2016), *Information Visualization* and *International Journal of Cartography*.

Natalia Andrienko is a lead scientist responsible for visual analytics research at Fraunhofer Institute for Intelligent Analysis and Information Systems and part-time professor at City University London. Results of her research have been published in two monographs "*Exploratory Analysis of Spatial and Temporal Data: a Systematic Approach*" (Springer 2006) and "*Visual Analytics of Movement*" (Springer 2013). Natalia Andrienko is an associate editor of *IEEE Transactions on Visualization and Computer Graphics*.